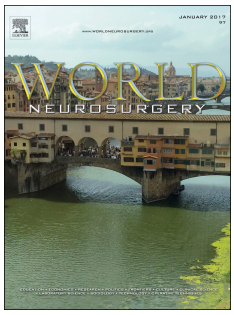
Journal Pre-proof



Technical Note: A Guide to Annotation of Neurosurgical Intraoperative Video for Machine Learning Analysis and Computer Vision

Dhiraj J. Pangal, BS, Guillaume Kugener, MEng, Shane Shahrestani, MS, Frank Attenello, MD, Gabriel Zada, MD, MS, Daniel A. Donoho, MD

PII: S1878-8750(21)00390-9

DOI: https://doi.org/10.1016/j.wneu.2021.03.022

Reference: WNEU 16941

To appear in: *World Neurosurgery*

Received Date: 11 January 2021

Revised Date: 2 March 2021

Accepted Date: 3 March 2021

Please cite this article as: Pangal DJ, Kugener G, Shahrestani S, Attenello F, Zada G, Donoho DA, Technical Note: A Guide to Annotation of Neurosurgical Intraoperative Video for Machine Learning Analysis and Computer Vision, *World Neurosurgery* (2021), doi: https://doi.org/10.1016/j.wneu.2021.03.022.

Author Credit Statement

Conceptualization: DJP, GK, GZ, DAD
Data curation: DJP, GK, GZ, DAD
Formal analysis: DJP, GK, GZ, DAD
Investigation: DJP, GK, GZ, DAD
Methodology: DJP, GK, GZ, DAD
Project administration: GZ, DAD
Resources: GZ, DAD
Software: DJP, GK
Supervision: GZ, DAD
Roles/Writing - original draft: DJP, GK, GZ, DAD
Writing - review and editing: DJP, GK, SS, FA, GZ, DAD

Technical Note: A Guide to Annotation of Neurosurgical Intraoperative Video for Machine Learning Analysis and Computer Vision

Dhiraj J Pangal BS[1], Guillaume Kugener MEng[1], Shane Shahrestani MS[1,2], Frank Attenello MD[1], Gabriel Zada MD, MS[1], Daniel A Donoho MD[1]

[1]Department of Neurosurgery, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[2]Department of Medical Engineering, California Institute of Technology, Pasadena, CA USA

Corresponding Author:
Dhiraj Pangal, BS
Department of Neurological Surgery
1200 North State Street, Suite 3300
Los Angeles, California 90033
Ph: 323-226-7421
Fax: 323-226-7833
pangal@usc.edu

Keywords: artificial intelligence, machine learning, computer vision, intraoperative video
Short Title: Annotating Surgical Video for Machine Learning
Previous Submissions: WN-20-7992
Previous Presentations: n/a
Word Count: 2183
Abstract Count: 250

Technical Note: A Guide to Annotation of Neurosurgical Intraoperative Video for Machine Learning Analysis and Computer Vision

Abstract

Objective: Computer vision (CV) is a subset of artificial intelligence which performs computations on image or video data, permitting the quantitative analysis of visual information. Common CV tasks that may be relevant to surgeons include image classification, object detection and tracking, and extraction of higher order features. Despite the potential applications of CV to intraoperative video, however, few surgeons describe the use of CV. A primary roadblock in implementing CV is the lack of a clear workflow to create an intraoperative video dataset to which CV can be applied. We report general principles for creating usable surgical video datasets and the result of their applications.

Methods: Video annotations from cadaveric endoscopic endonasal skull base simulations (n=20 trials of 1-5 min, size = 8GB) were reviewed by 2 researcher-annotators. An internal, retrospective analysis of workflow for development of the intraoperative video annotations was performed to identify guiding practices.

Results: Approximately 34,000 frames of surgical video were annotated**.** Key considerations in developing annotation workflows include: 1) Overcoming software and personnel constraints, 2) Ensuring adequate storage and access infrastructure 3) Optimization and standardization of annotation protocol, and 4) Operationalizing annotated data. Potential tools for use include CVAT and Vott: open-sourced annotation software allowing for local video storage, easy setup, and the use of interpolation.

Conclusion: CV techniques can be applied to surgical video, but challenges for novice users may limit adoption. We outline principles in annotation workflow that can mitigate initial challenges groups may have when converting raw video into useable, annotated datasets.

Background

The use of computational algorithms, particularly convolutional neural networks to analyze images or video is referred to as "computer vision" [1,2] (CV). There are many common CV tasks, including image classification, object detection, image segmentation, object tracking and the extraction of higher order features, and CV technology has been heavily utilized in the fields of autonomous vehicles, agriculture, and surveillance amongst others[3,4]. Surgeons have the potential to generate large quantities of visual data since the critical components of most cranial surgeries and many minimal-access spine surgeries are viewed using a device with camera capabilities. With many endoscopic and microscopic neurosurgical cases being recorded, hospitals and surgeons can potentially generate hundreds of hours of underutilized surgical video. Outside of neurosurgery, surgeons have attempted to analyze these large volumes of surgical video using CV , with the goal of eventually using these videos to predict patient outcomes, establish best practices, or assist in surgical training[5–8]. While new, groups outside of neurosurgery have found success in using CV to quantify surgical technique. Within urologic surgery, CV analysis of intraoperative video has been used to automatically identify surgical gestures[9]. Within laparoscopic surgery, CV techniques have similarly been successful at detecting the phase of surgery based on the detected presence of tools[10–12]. However within neurosurgery, CV remains underutilized particularly in the context of intraoperative video analysis.

Many CV pipelines have common elements. The first step in developing a CV pipeline is to create a library (dataset) composed of individual images (frames) from surgeries, to which "annotations" are applied (termed "ground-truth" data). These annotations are overlays that outline selected tools anatomical structures, or stages of an operation on a frame-by-frame basis[2,13,14]. Ground-truth data are often generated manually, and sometimes requires expert assessment. CV algorithms can be trained on this ground-truth data, then tested on newly inputted images or video. Thus, the successful development of any CV algorithm is fundamentally dependent on the quality, size, and accuracy of these annotated video sets.

Publications describing CV algorithms for surgical video do not elaborate on specific annotation techniques[13,15,16]. As a result, neurosurgery researchers looking to analyze their own video must essentially start anew- a painstaking and daunting task. This barrier prevents most groups' from analyzing their gigabytes of operative video available from surgical cases. In this manuscript, we outline fundamental considerations for annotation of surgical video developed through internal trial-and-error. While the literature does not currently have clear objective criteria to achieve the optimal methodology for annotation, it is our aim that the following protocols may serve as a starting point for future CV scholars.

Methods:

Surgical video from fresh tissue cadaver simulation of endoscopic endonasal skull base surgery cases (n=20 simulated trials between 1 and 5 min in length) were reviewed[17]. Neurosurgery and otolaryngology resident and attending surgeons were instructed to manage an iatrogenic internal carotid artery laceration (trial 1). Following an educational intervention by one of the senior authors, a second trial was conducted. Trials ended when hemostasis was achieved via muscle patch, or after five minutes of efforts (Trial Failure which resulted in simulated patient mortality). This protocol has been previously described in detail by our group[18].

Development of a surgical annotation methodology for neurosurgical video has been in progress at our institution since 2018. Raw intraoperative data was edited to include only relevant portions of trials. Videos were then annotated by the authors by using an annotation software to place bounding boxes around each surgical tool in frame. Following the development of our dataset, a retrospective internal analysis to identify 3-5 key elements for "developing annotated datasets from raw surgical video" was independently conducted by the lead authors (DJP, GK). These key points were then consolidated and categorized by consensus between the lead authors and senior authors (GZ, DAD). The subsequent sections serve to outline the details of annotated dataset development based on these consolidated key points.

Results

A total of approximately 34,000 frames of intraoperative video were annotated. The creation of surgical video annotation is time-consuming and tedious. Annotations were made frame by frame, and each tool/anatomical landmark (e.g. "grasper", "cottonoid", "artery", "dura") was outlined with a computer mouse; individual image labels (e.g. for staging an operation: "bleeding", "exploration", etc.) required assignment to each image (Figure 1). An alternative option is "segmenting"- where tools are traced with many individual line segments (versus contained within a box), providing a more specific identification of the tool (Figure 2). The decision for implementing annotations with bounding boxes vs segmentation, or other methods of annotation (e.g., identifying key points), would largely depend on the predetermined goals of the dataset. Bounding boxes can quickly and accurately gather information on tool (or anatomy) presence or absence. Segmentation, although more time consuming, may however be preferred if more information regarding angles of tools, tool location, or highlighting key anatomical relationships is required.

Typical endoscopic cameras operate at 30 frames-per-second, meaning one minute of surgical video has 1,800 annotatable frames, with full-length procedures potentially surpassing one million frames, each of which potentially requiring manual annota-

tion. Video datasets often have hundreds to a few thousand minutes of video to encompass the different images that may be presented, either due to anatomical variations, surgical technique differences, or videography characteristics (lighting, camera resolution, type of lens, etc).

As a result, converting raw surgical data into a robust ground-truth dataset is a daunting task. To efficiently mitigate these challenges requires four key considerations: 1) Annotation software and personnel to interface with the data, 2) Computing and storage infrastructure, 3) Developing an efficient and standard protocol and 4) Processing and Utilizing Annotated Data.

Annotation Software and Personnel

Software:

Many tools exist to label imaging and video-based datasets. These tools vary from free, lightweight and open-source, to cloud-based Software as a Service applications requiring monthly subscriptions. Surgical videos are often hours long and gigabytes in file size. Accordingly, lightweight and easy-to-install tools may lag or crash when dealing with such files. Web-based applications often require files to be uploaded to the application but are often similarly unable to handle large file sizes without crashing. Annotation tools that were built to handle large file sizes should be prioritized. To annotate, our group uses the open-source annotation tools CVAT[19] (Computer Vision Annotation Tool) and Vott[20]. Both tools take up minimal hard-drive space, are downloaded directly to the annotators' computer and do not require uploading raw video files prior to annotation. Lastly, simple features like the ability to copy-and-paste annotations between frames are instrumental in increasing efficiency but are not universally found on all software programs.

Personnel:

Due to the time-consuming nature of annotations, labs often outsource this work to medical students, undergraduates or other non-experts [21–23]. In our experience, an efficient user with the appropriate clinical knowledge could annotate about one frame per minute. Additionally, the implementation of a hierarchy with varying administrative privileges and annotator "supervisors" who can ensure consistency between users and provide quality-control. We suggest identifying platforms that facilitate collaboration and that include features allowing for easy cross-user label validation.

Alternatively, there are services such as Amazon Mechanical Turk© (MTurk) or Fiverr® which aim to expedite the process through paid professional annotators. While groups have developed high quality annotations from third-party services, costs of $1 for 10 images (0.3 seconds of video at 30 fps)[21], accelerate costs beyond what may be feasible for a pilot-study or proof of concept. However, the advantage of these services is it allows for additional annotations to be completed at scale- an advantage for groups with a more dedicated machine learning goal.

Infrastructure

Security of any personal health information (PHI) and accessibility for staff are the two main concerns with regards to video storage and annotation infrastructure. For raw-video storage, we recommend de-identifying videos before uploading to a HIPAA secure third-party client. Even if individual videos are primarily de-identified, the uploading of PHI onto private, third-party servers pose several ethical and legal questions that researchers and Institutional Review Boards must consider. For smaller labs with one or two annotators only, the use of cloud service providers (Google; Amazon Web Services) (Mountain View, CA; Seattle, WA) may be considered.

These data-privacy considerations must also be taken into account when discussing the chosen annotation software. Labeling platform needs to accept "local" files (i.e. files stored on a local machine or on a local server space owned by the institution) and we advise against using platforms that require the upload of raw data to 3rd party server

spaces as these may not be HIPAA compliant. These considerations must especially be taken into account when utilizing crowd-sourcing or MTurks, though HIPAA compliant options are available[24]. Of note, individual institutions may prohibit the use of these crowdsourcing platforms due to intellectual property concerns and/or concerns with vetting of annotators.

Developing an Efficient and Standardized Annotation Protocol

Maximizing annotation efficiency is integral when working with datasets with hundreds of thousands to millions of images. Here we outline key processes that have increased our internal efficiency.

The primary roadblock in annotation is the tediousness of outlining thousands of images. A potential solution is through the process of interpolation[22,25]. With interpolation, the user outlines an object, and adjusts the outline a set number of frames later. The software then smooths out the path between the two outlines, allowing 5-10 frames to be annotated using 2 keystrokes (Video 1). This improves the efficiency of annotation, as only minor corrections to the interpolation must be made. In our experience, a user can complete a minute of annotation in under an hour with interpolation, a thirty-fold improvement from annotating "from scratch", as determined by a two-user internal analysis (Table 1). Interpolation is a key consideration when approaching the annotation of videos within neurosurgery, where procedures are confined to small spaces with fine-controlled movements and may be hours long.

An additional strategy to reduce the number of frames for annotation is "downsampling". Endoscopes currently record video at 30 frames per second. Annotating only 10 or even 1 frame out of the thirty recorded can cut the number of frames annotated by orders of magnitude without losing significant power in computer models. This is a result of subsequent frames typically providing little novel visual information and decreasing the marginal utility of consecutive frames. This approach has been previously used, and may be particularly useful in endoscopic neurosurgery where tools

may be stationary for a large portion of time[26]. However, this strategy may limit the utility of future predictive models that rely on the temporal relationships between tools.

Interpolation and downsampling can be used together effectively to accelerate annotation efforts. By downsampling video (e.g. from 30fps to 1fps), the total number of frames needed to annotate is greatly reduced. By then also using interpolation, annotators can take advantage of the naturally smooth movement of tools and limit the number of keystrokes needed to annotate those frames, particularly when there the number of tools in view remain constant.

Other strategies have been used by groups to efficiently annotate video, such as using regression heatmaps to delineate objects, or utilizing deep neural networks to automatically annotate objects in frame[27]. These weakly supervised or unsupervised techniques are growing in popularity but were not utilized in our dataset development and further research is required for their implementation in a surgical database.

Standardization:

We recommend developing a *gold standard* annotation, where a supervisor can exemplify appropriate borders and clear identification of each tool/object in the frame. For interpolation and/or boxing, a few additional considerations must be made. For one, the rate of interpolation can be controlled by requiring annotators to re-adjust an annotation every "X" number of frames (step-size). A larger step-size may increase the volume of data annotated (useful for Deep Learning methods), whereas lowering the step-size will likely lead to an increase in quality (more useful for conventional neural networks). Additionally, guidelines must be provided to annotators that outline whether annotations should prioritize encompassing all of an object, a salient feature only, or limiting background in an annotation. Lastly, supervisors must iteratively decide what tools, anatomical landmarks, phases etc. need to be annotated, and update centralized instructions accordingly.

Operationalization of Annotated Data

Once annotated, the dataset can then be analyzed. However, there are several considerations that researchers need to make when storing and processing this data. Many file formats, such as COCO, PASCAL VOC, and YOLO, have been used to store individual image-level annotations for large, labelled image datasets[28–30]. Some of the file formats have publicly available development toolkits that allow users to interact with the information stored in the annotations at a basic level. However, for more complex analysis of information of these tools, users are left to expand on these toolkits with their own functions and methods or to write their own functionality from scratch. As such, the lack of tools available for researchers to begin analyzing their dataset can prove a significant hurdle, particularly for groups with limited computational expertise. Therefore, in the process of analyzing our own datasets, we are creating a repository containing scripts and classes we are continuously expanding on to perform critical tasks in our analysis, such as overlay annotations on images, extract surgical tool positions, and engineer new features based on the annotations.

Future Directions

The field of medical or computer vision, is a rapidly expanding field whose applications are far reaching. Within neurosurgery, the use of intraoperative image analysis could in theory help guide a surgeon to safely address anatomical boundaries. Analyzing surgical technique across institutions could aid in development of gold standard techniques. Finally, providing trainees with an objective review of their recent surgical cases would allow for reflection and improvement even without direct supervision. As the field of neurosurgical computer vision advances, similar advances in dataset development and annotation are needed in order to lower the barrier for entry for many surgeons who have the data but lack the infrastructure to conduct machine learning analyses.

Conclusions

The development of an annotated dataset of neurosurgical video that is appropriate for CV analysis can be a time-consuming and tedious process. While many surgeons have hundreds of hours of their own surgeries available to them, the inability to quickly and efficiently annotate these videos preclude their use at scale within the rapidly growing field of medical machine learning and CV. In this manuscript, we outline a protocol and considerations for groups looking to annotate their own surgical videos which has improved our efficiency. The potential for CV to augment surgical training and outcomes is clear, and the development of a gold-standard protocol for video annotation is a strong first step in achieving these goals.

FIGURE CAPTIONS

Figure 1. Bounded boxes outlining grasper, suction, cottonoid and string. Bounded boxes can be used for interpolation

Figure 2. Segmented tool outlines. Purple outline (grasper) is composed of many individual line segments (purple dots). Also shown are suction (blue outline) and cottonoid (yellow outline)

VIDEO CAPTION

Video 1: Example of interpolation being used to draw bounded boxes around surgical instruments

References

1. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. *Ann Surg*. 2018;268(1):70-76. doi:10.1097/SLA.0000000000002693

2. Ward TM, Hashimoto DA, Ban Y, et al. Automated operative phase identification in peroral endoscopic myotomy. *Surg Endosc*. Published online July 27, 2020. doi:10.1007/s00464-020-07833-9

3. Sedaghat-Pisheh H, Rivera AR, Biaz S, Chapman R. Collision avoidance algorithms for unmanned aerial vehicles using computer vision. *J Comput Sci Coll*. 2017;33(2):191-197.

4. Patrício DI, Rieder R. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Computers and Electronics in Agriculture*. 2018;153:69-81. doi:10.1016/j.compag.2018.08.001

5. Hung AJ, Chen J, Gill IS. Automated Performance Metrics and Machine Learning Algorithms to Measure Surgeon Performance and Anticipate Clinical Outcomes in Robotic Surgery. *JAMA Surg*. 2018;153(8):770-771. doi:10.1001/jamasurg.2018.1512

6. Hung AJ, Chen J, Ghodoussipour S, et al. A deep-learning model using automated performance metrics and clinical features to predict urinary continence recovery after robot-assisted radical prostatectomy. *BJU Int*. 2019;124(3):487-495. doi:10.1111/bju.14735

7. Makary MA. The Power of Video Recording: Taking Quality to the Next Level. *JAMA*. 2013;309(15):1591-1592. doi:10.1001/jama.2013.595

8. Makary MA, Xu T, Pawlik TM. Can video recording revolutionise medical quality? *BMJ*. Published online October 21, 2015:h5169. doi:10.1136/bmj.h5169

9. Luongo F, Hakim R, Nguyen JH, Anandkumar A, Hung AJ. Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery. *Surgery*. Published online September 26, 2020. doi:10.1016/j.surg.2020.08.016

10. Volkov M, Hashimoto DA, Rosman G, Meireles OR, Rus D. Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE; 2017:754-759. doi:10.1109/ICRA.2017.7989093

11. Hashimoto DAM, Rosman G, Witkowski ERM, et al. Computer Vision Analysis of Intraoperative Video: Automated Recognition of Operative Steps in Laparoscopic Sleeve Gastrectomy. *Annals of Surgery*. 2019;270(3):414-421. doi:10.1097/SLA.0000000000003460

12. Winkler-Schwartz A, Yilmaz R, Mirchi N, et al. Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation. *JAMA Netw Open*. 2019;2(8):e198363-e198363. doi:10.1001/jamanetworkopen.2019.8363

13. Funke I, Mees ST, Weitz J, Speidel S. Video-based surgical skill assessment using 3D convolutional neural networks. *Int J CARS*. 2019;14(7):1217-1225. doi:10.1007/s11548-019-01995-1

14. Singh A, Haque A, Alahi A, et al. Automatic detection of hand hygiene using computer vision technology. *J Am Med Inform Assoc*. Published online July 26, 2020. doi:10.1093/jamia/ocaa115

15. Hashimoto DA, Rosman G, Rus D, Meireles OR. Surgical Video in the Age of Big Data. *Ann Surg*. 2018;268(6):e47-e48. doi:10.1097/SLA.0000000000002493

16. Quellec G, Charrière K, Lamard M, et al. Real-time recognition of surgical tasks in eye surgery videos. *Med Image Anal*. 2014;18(3):579-590. doi:10.1016/j.media.2014.02.007

17. Kugener G, Pangal D, Cardinal T, Collet C, Zhu Y. Prediction of Neurosurgical Hemorrhage Control and Instrument Detection Using Deep Learning. *International Conference on Information Processing in Medical Imaging 2021 (Submitted)*.

18. Pham M, Kale A, Marquez Y, et al. A Perfusion-based Human Cadaveric Model for Management of Carotid Artery Injury during Endoscopic Endonasal Skull Base Surgery. *J Neurol Surg B*. 2014;75(5):309-313. doi:10.1055/s-0034-1372470

19. Sekachev B, Nikita M, Andrey Z. *Computer Vision Annotation Tool: A Universal Approach to Data Annotation*. https://software.intel.com/en-us/articles/computer-vision-annotation-tool-a-universal-approach-to-data-annotation

20. *Microsoft/VoTT*. Microsoft; 2020. Accessed July 7, 2020. https://github.com/microsoft/VoTT

21. Zhou N, Siegel ZD, Zarecor S, et al. Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning. *PLoS Comput Biol*. 2018;14(7). doi:10.1371/journal.pcbi.1006337

22. Zhang L, Chen X-Q, Kong X-Y, Huang H. Geodesic Video Stabilization in Transformation Space. *IEEE Trans Image Process*. 2017;26(5):2219-2229. doi:10.1109/TIP.2017.2676354

23. Vondrick C, Patterson D, Ramanan D. Efficiently Scaling up Crowdsourced Video Annotation: A Set of Best Practices for High Quality, Economical Video Labeling. *Int J Comput Vis*. 2013;101(1):184-204. doi:10.1007/s11263-012-0564-1

24. Architecting for HIPAA Security and Compliance on Amazon Web Services. :56.

25. Rufenacht D, Taubman D. HEVC-EPIC: Fast Optical Flow Estimation from Coded Video via Edge-Preserving Interpolation. *IEEE Trans Image Process*. 2018;27(6):3100-3113. doi:10.1109/TIP.2018.2813090

26. Jin A, Yeung S, Jopling J, et al. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE; 2018:691-699. doi:10.1109/WACV.2018.00081

27. Payer C, Štern D, Bischof H, Urschler M. Regressing Heatmaps for Multiple Landmark Localization Using CNNs. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Vol 9901. Lecture Notes in Computer Science. Springer International Publishing; 2016:230-238. doi:10.1007/978-3-319-46723-8_27

28. Lin T-Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context. *arXiv:14050312 [cs]*. Published online February 20, 2015. Accessed March 17, 2020. http://arxiv.org/abs/1405.0312

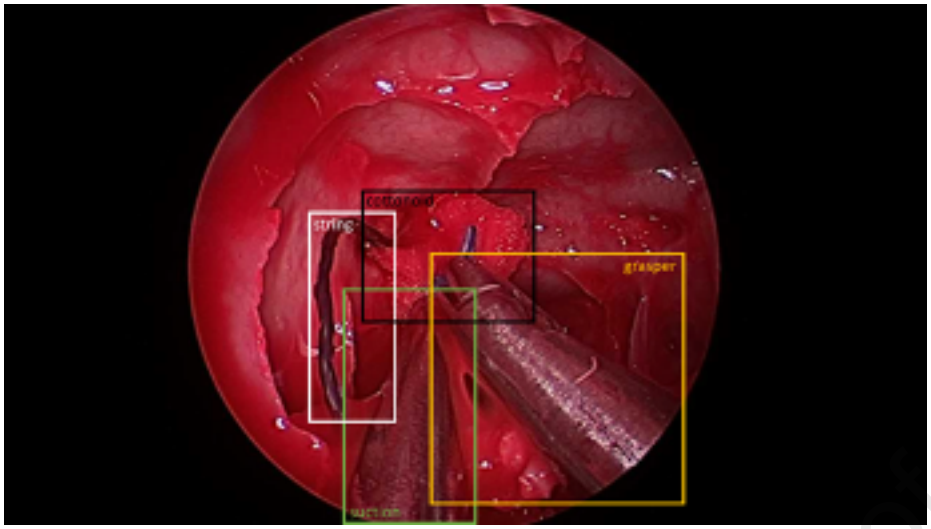
29. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2016:779-788. doi:10.1109/CVPR.2016.91

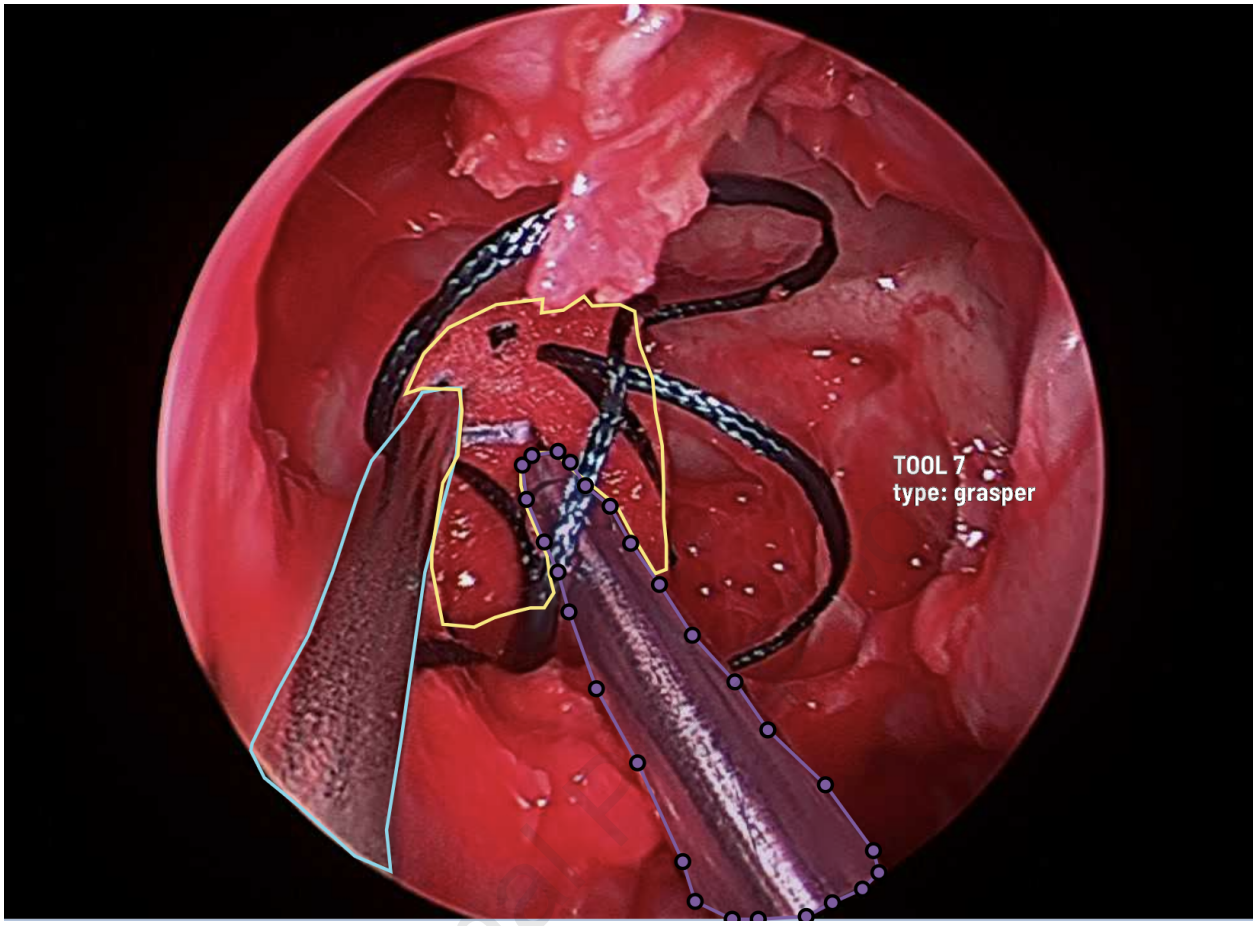
30. Everingham M, Van-Gool L, Williams C, Winn J, Zisserman A. *The {PASCAL} {V}isual {O}bject {C}lasses {C}hallenge 2008 {(VOC2008)} {R}esults*. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html

TABLES

	User 1	User 2
Boxed, Interpolated	550 frames	110 frames
Boxed, Non-Interpolated	26 frames	28 frames
Segmented, Non-Interpolated	19 frames	20 frames

Table 1: Annotation rate (timed, 15 minute interval) between two users for boxed annotation of tools with interpolation, boxed annotation without interpolation, and segmental outlining of tools without interpolation. Annotation software did not permit interpolated segmental outlining.





Abbreviations

CV: Computer Vision
CVAT: Computer Vision Annotation Tool
COCO: “Common Objects in Context”
PASCAL: Pattern Analysis, Statistical Modeling and Computational Learning
VOC: Visual Object Classes
XML: Extensible Markup Language
YOLO: “You Only Look Once” object detection system